GENOMICS 1, 103–106 (1987)

EDITORIAL

Toward a Complete Map of the Human Genome

There are many ways to portray the linear organization of the elements of the chromosomes. There are many types of chromosome maps just as there are many ways to represent the geographic map. In general, the human genome can be represented in physical maps or in genetic maps.

The Physical Map

The photograph of the high-resolution banded karyotype and the idiogram derived therefrom is a basic physical map and the map at lowest rèsolution, perhaps 1000 elements at the most. It is comparable to mapping by aerial or satellite photography. The map positioning expressed genes in relation to chromosome bands is another type of physical map; most of the map created by interspecies somatic cell hybridization and by *in situ* chromosomal hybridization is of this type (for mapping, however, molecular genetic methods make it unnecessary that the segment of DNA be expressed). At least 1200 expressed genes have been placed on the physical map, i.e., assigned to specific chromosomes and in many instances to specific bands.

A map of overlapping cloned DNA segments is a third type of physical map—a "contig" map; this can be related to the band map and expressed genes can be related to contigs. The contigs can be of sizes varying from 20 to several thousand kilobases, depending on the method of cloning. The internal structure of each contig can be specified by the pattern of cutting by various restriction endonucleases—the restriction map.

The nucleotide sequence is the ultimate physical map. The haploid set of chromosomes in the human contains about 3.0 billion base pairs; 6.0 billion base pairs are present in the diploid set (23 chromosome pairs).

Genetic Maps (Linkage Maps)

The primary genetic map depicts the location of gene loci (expressed DNA segments) in relation to each other. The intervals between loci are determined by the extent to which crossingover (recombination) occurs between the loci. Genetics is the science of biologic variation. Variation at the loci in question is essential to genetic mapping. In families, cosegregation of alternative forms of each of the two DNA segments called loci is the means by which physical proximity of the loci is recognized ("genetic linkage"); the extent to which cosegregation deviates from completeness is the measure of genetic distance between the loci. Thus, recombination can vary from virtually zero, if the loci are very-close, to 50%, if they are far apart on the same chromosome or located on different chromosomes. The amount of recombination (as percentages) can be expressed in centimorgans (cM). The occurrence of double crossingover between two loci located more widely from each other has the effect that the recombination fraction (and cM value derived therefrom) does not have a linear relationship to physical distance on the chromosome. Other complications in relating the genetic map to the physical map are (1) the occurrence of recombination "hotspots," i.e., inhomogeneity in the distribution of crossovers, and (2) differences in the frequency of crossingover in males and females, this being greater in females in most but not all parts of the genome. The total genetic length of the human genome is estimated to be something over 3000 cM.

The variation in DNA that is the basis of genetic mapping may be reflected in the phenotype, i.e., in disorders such as Huntington's disease or cystic fibrosis, or in immunologic, electrophoretic, or physiologic variation, such as ABO blood group, esterase D type, or colorblindness. On the other hand, the variation in DNA can be more directly detected by the pattern of cutting by restriction endonucleases (restriction fragment length polymorphisms, RFLPs, "riflips"). The latter variation can be used to create a reference map of the human genome which for maximal usefulness might have a RFLP marker located each 1 cM, on the average. Studies for creation of a "saturated" RFLP reference map can be done in "libraries" of the DNA from a limited number of "complete" families: a sibship of eight or more children plus the parents and four grandparents. Panels of such families are maintained, for example, in Paris by the Centre d'Étude du Polymorphisme Humain (CEPH) of Jean Dausset and in Salt Lake City by the unit of Ray White. DNA clones containing RFLPs can be prepared from individual chromosomes separated by recently developed methods. Thus, a "primary genetic map" of individual chromosomes can be pre-

0888-7543/87 $3.00

pared. VNTR (variable number of tandem repeats) recognized by certain restriction enzymes are yet another form of polymorphism highly valuable in creating the primary genetic map.

Like the map of expressed genes, the individual RFLP markers and the RFLP map as a whole can be related to the band map by somatic cell hybridization and *in situ* hybridization, and to the contig map, as well.

The cDNA Map

Another type of physical map is the cDNA map, also called exon map or mRNA map. For the cDNA map, complementary DNAs synthesized by reverse transcription from messenger RNA, regardless of whether the protein gene product of that mRNA is known or not, are assigned to chromosomal bands by a combination of somatic cell hybridization and *in situ* hybridization.

Why is the cDNA map particularly important in this stage of human genomics? The reason is that creation of the cDNA map is an excellent way to find the expressed parts of the genome, to find the structural genes that determine the amino acid sequences of proteins and therefore are of particular functional significance. This part of the genome, representing with its introns and immediate flanking regions about 5% of the total, is a front contender for early sequencing.

Mendelian Inheritance in Man (MIM) is an encyclopedia of expressed gene loci. No more than one entry per locus is wittingly created in MIM. Until rather recently MIM was based almost exclusively on mendelizing phenotypic variation. When various phenotypes were known to be the result of different mutations at a single locus, all were incorporated into a single entry. A classic example is the β-globin locus, mutation at which can cause cyanosis (β forms of hemoglobinopathic methemoglobinemia), erythremia (resulting from forms of the β-globin chain with increased oxygen affinity), or anemia, which itself can take various forms (e.g., β-thalassemia, sickle cell anemia, and Heinz-body anemia from unstable hemoglobin). In the past in a few instances when a protein was completely sequenced, a MIM entry was created for it even though no mendelian variation had been identified. In more recent times, when an expressed gene has been cloned and sequenced, it has been honored with a MIM entry, again regardless of whether mendelian variation is known.

The total number of expressed genes represented in the MIM encyclopedia (i.e., in its online continuously updated version, OMIM), according to the count of October 1, 1987, was 4257; only about half of these are asterisked (i.e., considered fully established). (See Fig. 1.) There are surely duplications (incorrect "splitting") and contrariwise there is certain to be "lumping" of phenotypes that represent more than one locus. Be that as it may, the number 4257 is far short of the estimated 50,000 to 100,000 structural genes in the human.

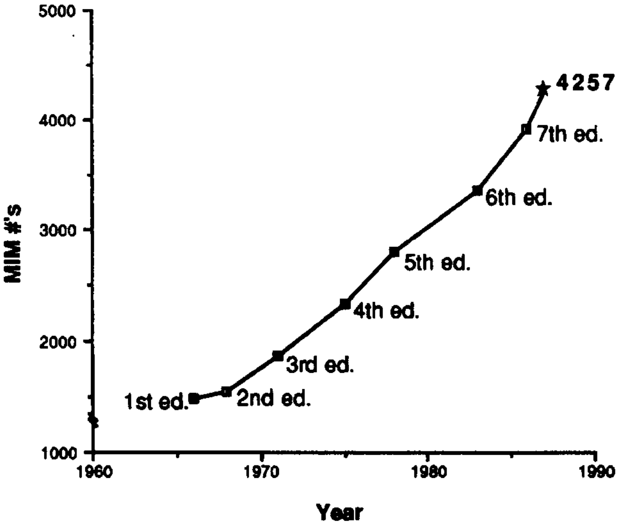

FIG. 1. Total number of entries in each edition of *Mendelian Inheritance in Man* (Johns Hopkins University Press) and, as of October 1, 1987 (star), in OMIM, the online, continuously updated version (Welch Medical Library, The Johns Hopkins University).

At HGM9 (the Paris Workshop on Human Gene Mapping) which concluded on September 11, 1987, it was announced that 1359 genes had been mapped to specific chromosomes and in most instances to specific regions of those chromosomes. A more accurate count of expressed genes may be 1215 (Fig. 2). But either number represents an impressive fraction of known genes. At the same time, however, it points up the potential usefulness of cDNA mapping in "finding the rest of the genes" of man and characterizing them.

Mapping/sequencing the human genome will help identify the rest of the expressed genes. When the full nucleotide sequence of the human is in hand, it should be possible to identify coding segments ("open reading frames") and to distinguish expressed genes from pseudogenes and exons from introns. But before that time, with techniques now available and easily capable of upscaling, the coding parts of the human genome can be identified and chromosomally mapped using cDNAs synthesized from mRNAs by reverse transcription. These cDNAs can be prepared from mRNAs, even those of low abundance, that are synthesized in differentiated tissues and at particular stages of development. Once the gene is cloned as cDNA it can be mapped by somatic cell hybridization and *in situ* hybridization. It can provide the basis for "pulling out" the corresponding genomic sequences (including the introns and flanking sequences with

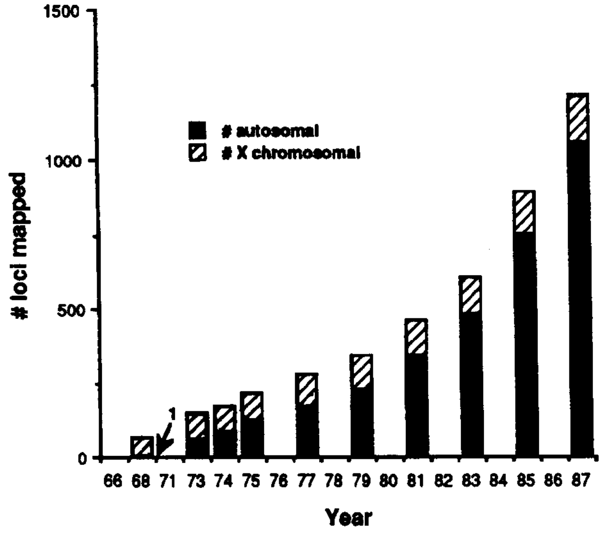

FIG. 2. Number of loci (*expressed* genes)-specific chromosomes at the time of each of the nine international workshops on human gene mapping, 1973–1987 (proceedings of each workshop published in *Cytogenetics and Cell Genetics* and by March of Dimes Birth Defects Foundation as part of its *Birth Defects: Original Article Series*). Also indicated is the number of recognized X-linked loci in 1968 at the time of the first mapping of a specific locus to an autosome: Duffy blood group to chromosome 1. Colorblindness was the gene first recognized as located on the X chromosome, in 1911.

promoters and enhancers). It can provide a useful starting point for sequencing. The derivation of genomic clones may facilitate demonstrations of RFLPs in introns and flanking sequences, useful in localizing the cDNAs on the genetic map—and a contribution to development of the saturated RFLP reference map. Many of the cDNAs are likely to represent complex genes with very large introns, multiple promoters, and alternative splicing and polyadenylation. The approach to these genes through the mRNA should help unravel this complexity.

Even if the functions of many of the individual genes were not known, a "saturated" mRNA map would be of tremendous value. By definition, it is of great biologic interest because this is where the action is. The mapped cDNAs can serve as "candidate genes": Does a given disorder of unknown genetic basis map to the same locale as does the cDNA?

It is estimated that 30–50% of the human genes are expressed only or mainly in the brain. (Estimates of the number of different mRNAs in the human brain run as high as 150,000. It is likely, however, that these represent a considerably smaller number of genes because of differential splicing that creates several peptides from a single gene, differential polyadenylation and recombinational events that create a large number of different mRNAs from a few genes, and other special mechanisms.) Imagine the assistance to the neurobiologist that could come from availability of a saturated map of brain cDNAs. In general, the construction of the cDNA map (with the cataloguing of the cDNAs that this will entail) will provide the groundwork for biologic studies that presently cannot proceed until the same work is done piecemeal and inefficiently before getting on with the biomedically important research.

Correlation of the Physical and Genetic Maps

A cluster of genes mapped by genetic means can be positioned in a particular band or set of bands by somatic cell hybridization or *in situ* hybridization of one or more members of the "linkage group." By macrorestriction mapping, with rare cutter endonucleases and pulsed field gel electrophoresis, closely situated genes can be rather precisely positioned relative to each other through hybridization of clones to large fragments. The orientation of the cluster in relation to the centromere may be determinable by centromere mapping using a centromere-related heteromorphism or centromeric DNA probe as a linkage marker. The orientation may also be achieved by the study of deletions or chromosomal rearrangements, particularly reciprocal translocations involving the region. In making physical–genetic correlations, the nonrandom distribution of crossingover becomes evident: suppression of crossingover near the centromere and in some other segments, recombination "hotspots" in yet other regions, and more crossingover toward the telomers.

The Morbid Map of the Human Genome

A physical map that represents the location of genes containing disease-producing mutations—the morbid map—is constructed in part from the physical map of the protein gene products that are defective in the several disorders; in part from genetic mapping of the disorder itself by linkage to markers that in turn have been physically mapped; and in part from physical mapping of the disorder itself, through deletion mapping, for example. (The mapping of the intragenic lesion in many mendelian disorders and neoplasms (somatic cell genetic disorders) is proceeding apace with the application of molecular genetic methods to demonstration of microdeletions, nucleotide substitutions, insertions, and other changes.)

How to Proceed with Mapping the Human Genome

As implied near the outset, the nucleotide sequence is the ultimate map, the chromosome map of highest resolution. It seems likely that knowledge of the complete sequence would be useful. It will be achieved most efficiently and meaningfully through mapping efforts begun at lower resolution. The catalogued and mapped cDNAs are a logical place to begin sequencing. Large DNA segments (contigs), well characterized by restriction patterns and mapped to specific chromosomal sites, will be the raw material for large-scale sequencing.

With relatively modest modification and enhancement, present technology is capable of accelerated mapping of RFLPs and cDNAs. This could be implemented immediately, with a coordinated effort for the more routine technical aspects and for information handling, including analysis (a very important part of the undertaking). Sequencing of the cDNAs and their genomic counterparts—and of the clones defining RFLPs—can proceed immediately. Correlation of the cDNA map with the RFLP map and other genetic maps should also proceed at once. Meanwhile, methods for cloning large DNA segments and facilitated sequencing technology should be worked out and ready to apply by the time the basic cDNA and RFLP maps have been "saturated."

We like goals! We need goals! How about the year 2000 for complete mapping/sequencing of the human genome? But it is doubtful that the goal will be achieved before the 21st century unless there are incremental ("add-on") resources for the undertaking and specific organization for management and coordination of a special effort.

Victor A. McKusick
Frank H. Ruddle